TEC-0088

Research in Information Science & Technology: Machine Vision

P. Allen
J. Kender
S. Nayar
T. Boult

The Trustees of Columbia University
in the City of New York
Department of Computer Science
500 West 120th Street
New York, NY 10027

August 1998

Approved for public release; distribution is unlimited.

U.S. Army Corps of Engineers
Topographic Engineering Center
7701 Telegraph Road
Alexandria, Virginia 22315-3864

US Army Corps
of Engineers
Topographic
Engineering Center

T
E
C

U.S. ARMY TOPOGRAPHIC ENGINEERING CENTER
TEC

Destroy this report when no longer needed.
Do not return it to the originator.

The findings in this report are not to be construed as an official Department of the Army position unless so designated by other authorized documents.

The citation in this report of trade names of commercially available products does not constitute official endorsement or approval of the use of such products.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. AGENCY USE ONLY *(Leave blank)*	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	August 1998	Technical April 1994 - April 1995

4. TITLE AND SUBTITLE

Research in Information Science & Technology:
Machine Vision

5. FUNDING NUMBERS

DACA76-92-C-0007

6. AUTHOR(S)

P. Allen, J. Kender, S. Nayar, T. Boult

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

The Trustees of Columbia University, City of New York
Department of Computer Science
500 West 120th Street
New York, NY 10027

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)

U.S. Army Topographic Engineering Center
7701 Telegraph Road
Alexandria, VA 22315-3864

19. SPONSORING / MONITORING AGENCY REPORT NUMBER

TEC-0088

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION / AVAILABILITY STATEMENT

Approved for public release; distribution is unlimited.

12b. DISTRIBUTION CODE

13. ABSTRACT *(Maximum 200 words)*

Machine Vision is fast becoming a key technology, and advances in machine vision are occurring along several fronts. This report outlines the progress being made at Columbia University in developing new machine vision algorithms and applications, and the associated technology transfer. Specific results include a new model of Lambertian reflectance, methods for recovery of shape from specularity, integrating color and polizaration for shape recovery, visual learning of appearance for fast object recognition, automated 3-D model acquisition from range imagery, new methods for modeling deformable objects, deriving shape from shadow information, methods to control robotic hands with vision, new approaches to sensor planning and placement, generating spatial language descriptions from imagery, and vision algorithms to recognize hand gestures.

14. SUBJECT TERMS

Object Recognition, Physics-Based Vision, Robotic Vision, Visual Integration

15. NUMBER OF PAGES

29

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
UNCLASSIFIED	UNCLASSIFIED	UNCLASSIFIED	UNLIMITED

Contents

List of Figures

PREFACE

This report was prepared under Contract DACA76-92-C-0007 for the U.S. Army Topographic Engineering Center, Alexandria, Virginia 22315-3864 by Columbia University, New York, NY 10027. The Contracting Officer's Representative was Ms. Lauretta Williams.

1 INTRODUCTION

This is the annual report for Darpa Contract DACA76-92-C-007, covering the period 3 April 1994 to 2 April 1995. In this report, we will highlight our progress and accomplishments across the full spectrum of research in machine vision, including physical models for image understanding, object recognition and modeling using vision, robotic vision, and integration of vision with hand gestures and language. More detailed results can be found in the papers listed in the bibliography and the most recent Advanced Research Projects Agency (ARPA) Image Understanding Workshop proceedings.

2 PHYSICAL MODELS FOR IMAGE UNDERSTANDING

2.1 Seeing Beyond Lambert's Law

Reflectance may be viewed as the first physical mechanism in the process of visual perception by man or machine. Hence, accurate reflectance models are key to the advancement of image understanding. In last year's report, we reported the development of a new comprehensive model for diffuse reflectance. This model was demonstrated to be an extensive generalization of the popular Lambert's model, and has far reaching implications for machine vision, visual psychophysics, computer graphics, and remote sensing. It shows that diffuse reflection can deviate significantly from Lambert's Law as the macroscopic roughness of a surface increases, causing the surface to appear brightest in the direction of the light source rather than the surface normal direction or the specular direction. We have conducted experiments on several natural and man-made surfaces such as clay, cloth, plaster, and wood, to demonstrate that the model subsumes a wide spectrum of real-world surfaces. In all cases, the accuracy of the model was found to be consistently high. In addition, the model was recently used for realistic graphics rendering. These results were presented at the 1994 SIGGRAPH conference held in Orlando in July, 1994 [20]. The implications of the model for computational vision were published in the International Journal of Computer Vision in March, 1995 [21].

The above reflectance model has lead us to an interesting observation. It predicts that very rough objects, when illuminated from close to the viewing direction, generate images of nearly constant brightness. In other words, the object, irrespective of its shape (be it polyhedral or smoothly curved), will produce just a silhouette, devoid of any brightness variations or shading! This visual effect is illustrated in Figure 1 which shows two cubes of the same dimensions, made from *exactly* the same material (clay), and illuminated from the same direction. The only factor causing the two cubes to appear different is their surface roughness. The cube on the left is fairly smooth and behaves very much like a Lambertian object with its three visible faces producing different brightness values. In contrast, the high roughness of the cube on the right causes all points on its surface to produce more or less the same brightness in the viewing direction, producing a silhouette of the object without any clearly discernible edges. Therefore, in the case of very rough diffuse objects illuminated from close to the viewing direction, shape from shading, by man or machine, becomes impossible! The implications of these results were reported in the journal *Science* in February, 1995 [19].

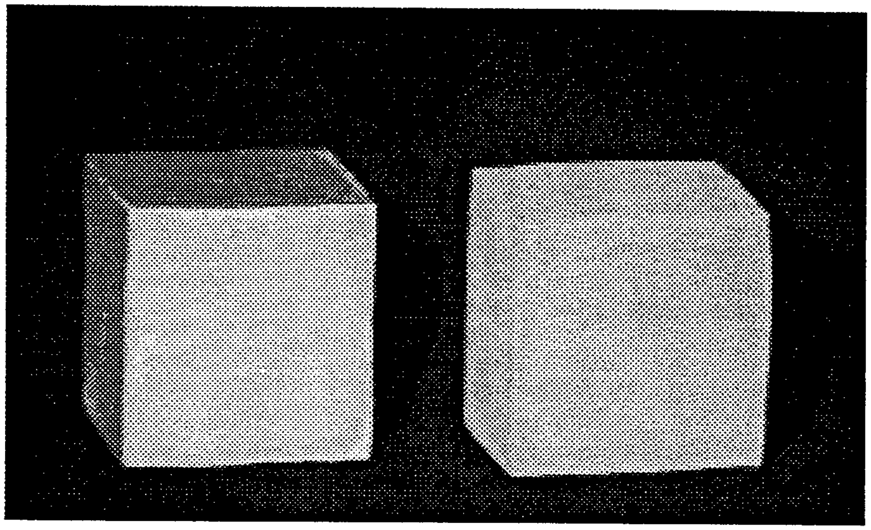


Figure 1: Two cubes of similar dimensions, made from exactly the same material (clay), and illuminated from the same direction (close to the viewing direction). The cube on the left is fairly smooth and behaves very much like a Lambertian object, while the one on the right has high roughness causes all points on its surface to produce more or less the same brightness; a silhouette without any clearly discernible edges. This effect is predicted by the developed reflectance model.

2.2 Recovery of Specular Surfaces

We have developed a theoretical framework for the perception of the three-dimensional geometry of specular (mirror-like) surfaces. Such surfaces are commonplace and the recovery of specular shapes has been to known to be a hard problem to solve. While it is not possible to compute specular structures from a single image, we have shown that a moving observer has sufficient image constraints to estimate shape unambiguously. When an observer moves in three-dimensional space, real scene features, such as surface markings, remain stationary with respect to the surfaces they belong to. In contrast, a virtual feature, which is the specular reflection of a real feature, travels on the surface. Based on the notion of caustics, a novel feature classification algorithm was developed that distinguishes real and virtual features from their image trajectories that result from observer motion. Next, using the support function representation of curves, a closed-form relation was derived between the image trajectory of a virtual feature and the geometry of the specular surface it travels on. We showed that a specular surface profile can be uniquely recovered by tracking just two unknown virtual features (see Figure 2). A publication on these results was awarded the David Marr Prize at the 1995 International Conference on Computer Vision held in Boston in June, 1995 [22].

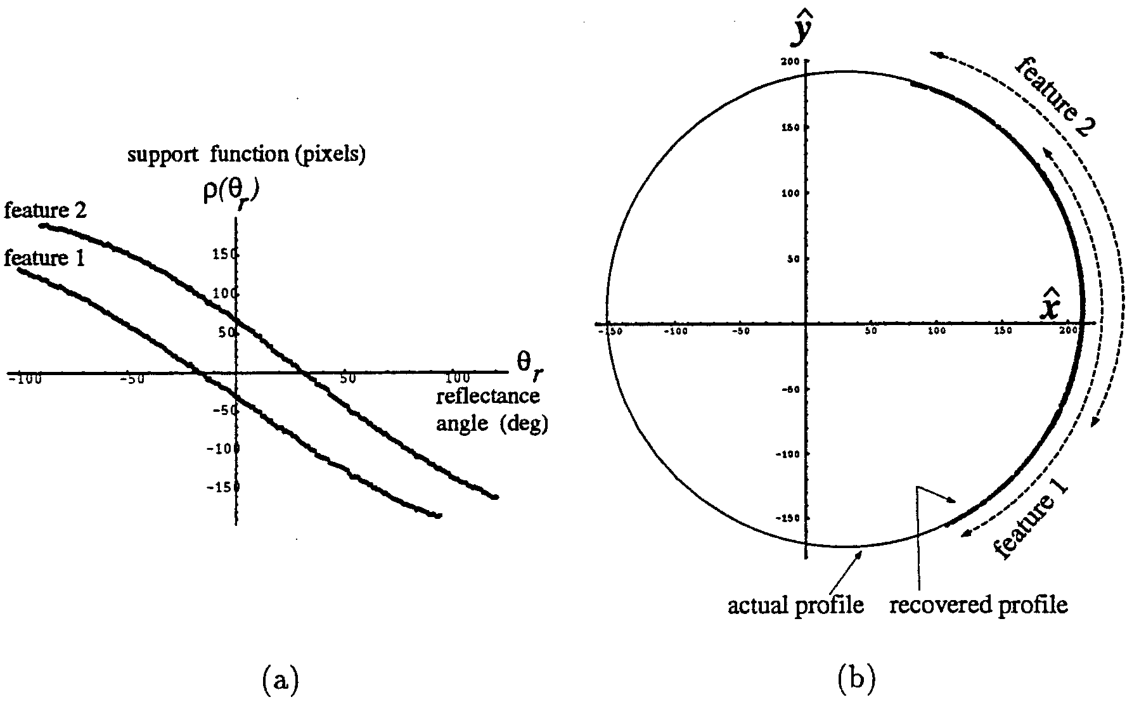


Figure 2: 2-D profile of a sphere recovered by tracking two unknown features. (a) Support functions of the two features computed from their image trajectories. (b) The recovered surface profile. The dots represent the computed profile and the solid line is the actual profile.

All area-based stereo algorithms are based on the implicit assumption that points in the scene are Lambertian in reflectance and hence corresponding points in stereo images have identical brightness values. However, specular reflection is highly viewpoint dependent. As a result, specularities can cause large intensity differences at corresponding points in stereo images. We have analyzed the physics of specular reflection and the geometry of stereopsis to arrive at an interesting relation between stereo vergence, surface roughness, and the likelihood of a correct stereo match. This result has led to a multiple-view stereo configuration that produces accurate depth maps in the presence of specular reflections. Several experiments were conducted on surfaces with varying reflectance properties to demonstrate the benefits of the proposed stereo configuration. These results were presented at the 1995 International Conference on Computer Vision held in Boston in June, 1995 [6].

2.3 Color, Polarization and Roughness

Continuing our work from last year, we finished our experimentation on the basic integration of color and polarization, providing for light source localization, color correction and highlight removal and their use in photometric stereo. The paper detailing the results of this work has been accepted to the Internation Journal of Computer Vision (IJCV) [18]. Working with two undergraduates we have made details of the experimentation available on the WWW at URL http://www.eecs.lehigh.edu/boult/POLAR.

Our previous work on polarization characteristics assumed "optically" smooth surfaces. It is natural to ask how the polarization characteristics vary as the surface deviates from "smooth". Again the issue of how to "model" roughness and the scale of roughness must be addressed. Because the inspection of surface finish is an important manufacturing application, we have decided to concentrate on micro-scale roughness, with surface variations between 10^{-4}m down to 10^{-9}m (Equivalent to "sanding" surfaces with emery cloth from grit 240 down to 4000.) Accurate, non-contact high-speed measurements of roughness at this level is also important in process-control as it can be used to indirectly detect/monitor wear of the machining and/or extrusion components in a manufacturing cell.

We could try and follow the approach of Nayar and Oren, building up a predictive model based on the analysis of a V-groove model of the roughness. However, because polarization is effected by both surface (specular) reflection and "diffuse" reflection, that analysis would be quite difficult. Further difficulties arise for roughness on the scales on the order of a wavelength $< 10^{-6}$m) because diffraction causes depolarization in a complex manner. Thus we have been pursuing roughness measurements via a "bulk" roughness model, or what has become known as an effective-medium-theory (EMT).

Numerous papers have studied the basic polarization affects of reflection from rough surfaces. The EMT model replaces usual Fresnel reflection coefficients $F_\perp$and $F_{||}$with "effective" Fresnel coefficients which are scaled as a "roughness" factor.

$$\begin{aligned}\hat{F}_\perp(\eta,\psi,) &= \rho(\psi,\frac{\sigma}{\lambda})F_\perp(\eta,\psi)\\ \hat{F}_{||}(\eta,\psi,) &= \rho(\psi,\frac{\sigma}{\lambda})F_{||}(\eta,\psi)\end{aligned}$$

where $F_\perp(\eta,\psi)$ and $F_{||}(\eta,\psi)$, are the Fresnel reflection coefficients of a smooth surface with index of refraction η and $\rho(\psi,\frac{\sigma}{\lambda})$ is a measure of roughness which for material with finite conductance is given by:

$$\rho(\psi,\sigma) = e^{\left(-4\pi cos(\psi),\frac{\sigma}{\lambda}\right)^2} \tag{1}$$

where ψ is the angle of incidence, λ is the wavelength of light, and σ is the standard deviation of the (assumed Gaussian) surface.

By considering the full functional form of the Fresnel coefficients, one can derive that at $\psi = 45°$ we have

$$\frac{\hat{F}_\perp^{\;2}}{\hat{F}_{||}} = e^{\left(-8\pi,\frac{\sigma}{\lambda}\right)^2} \tag{2}$$

independent of the material's index of refraction! This important quantity is called the specularity index of the surface. Thus, by approximating the specularity index, we can, given assumptions on the roughness model, infer the surface roughness. With a full ellipsometer, the components of effective Fresnel coefficients can be computed and the ratio in equation 2 directly computed. As in much of our past work with polarization-based vision, the goal of this research is to find effective approximations which can be computed using vision sensors. In particular, we seek techniques for computing ρ which do not require full ellipsometric techniques and are amenable to near-real time computation. Techniques we anticipate researching will include using multiple angles of incident and multiple colors (wavelengths). (To compute larger scale roughness we will need to consider polarization in the IR domain.)

The assumptions of the EMT model break down for roughness which is significantly above the wavelength of light used in the computation. Again, experimentation has shown the EMT based polarization measurements can still be repeatable and correlated to roughness, but less accurately predictive of the quantitative roughness. This previous work, using ellipsometers, has only used the specularly reflected component of light, and ignored the "diffuse" component. This assumption may have limited in the range of roughness for which the techniques could be applied, because for surfaces rougher than $10^{-6}m$, the signal to noise ratio (i.e. the ratio of specular reflection to diffuse reflection +real noise) becomes quite small. Using our combined specular/diffuse polarization model, we have begun studying the EMT model for small scale roughness. By using longer wavelengths imaged by near IR camera we expect that this can be extended to larger scale roughness, say 10^{-4} (depending on the material color or diffused albedo). Furthermore, by combining both specular and diffuse model we hope to reduce the accuracy requirements needed for existing levels of roughness.

While out primary goal is to determine quantitative surface roughness measurements, we are currently assuming a rather simple models of surface textures, Gaussian "bumps". However, real surfaces often have complex and anisotropic textures. Some issues of anisotropic textures and EMT models have been addressed in the literature, but the results are quite limited. While deriving quantitative measurements quickly becomes complex, it is clear that repeatable "polarimetric" information which is strongly correlated to the texture size and pattern, can be obtained. Therefore, we have begun investigations into using the SLAM package (see section 3.1) to directly "learn" roughness from appearance. As is often the case in visual learning, investigations into the proper representation of the input will be crucial. This has been the subject of our study to date.

3 OBJECT RECOGNITION AND MODELING USING VISION

3.1 Visual Learning and Recognition

Learning can often be viewed as the problem of determining a function that maps an input vector space to an output vector space. Training examples of such a mapping are used to

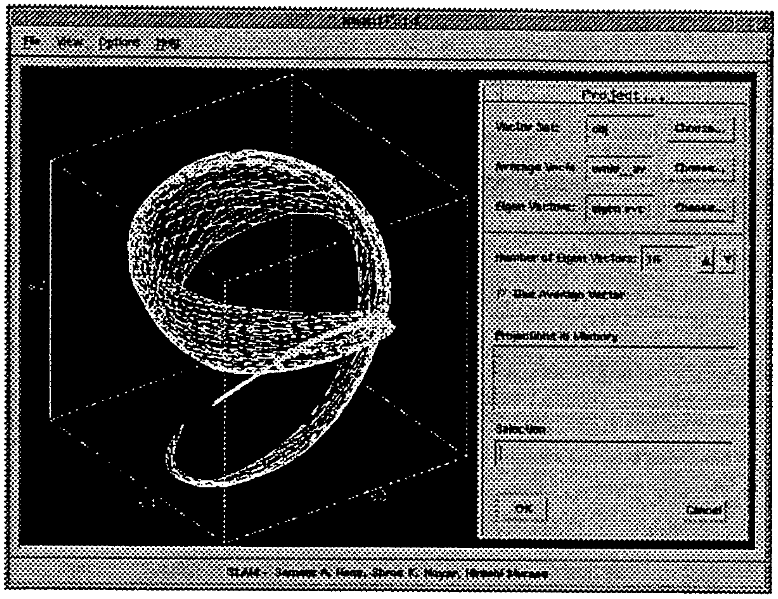


Figure 3: The SLAM software package has been developed as a general tool for appearance modeling and recognition problems in vision.

construct a continuous function that approximates the training data and generalizes it for intermediate instances. We have used generalized radial basis function (GRBF) networks to formulate the approximating function. We have introduced a novel method for constructing an optimal GRBF network for any given training data and an error bound using the integral wavelet transform. We have shown the optimality of the generated networks for the multi-dimensional mapping problem of object recognition and pose estimation. These results were presented at the 1995 International Conference on Computer Vision held in Boston in June, 1995 [13]. An extended version of this publication will appear in the Pattern Recognition Journal in 1996.

We have also developed a general algorithm for learning and recognizing visual appearance. We proposed the first parametrized model for visual appearance that can be automatically learned from two-dimensional images. The learning algorithm was used to develop a real-time color 3-D recognition system with 100 objects, real-time robot positioning and tracking, and illumination planning for object recognition in structured environments. Recognizing the scope and generality of our appearance matching framework, we developed a software library for appearance matching (SLAM) which includes all the modules needed for appearance learning and recognition (see Figure 3). This package has already been licensed to approximately 30 academic research laboratories around the world and 3 industrial sites in the U.S. (see section 6). In the last month, Columbia University has contacted approximately 100 machine vision companies. The response has been enthusiastic and we are in the process of working out license agreements with several companies. Our research on appearance matching can, therefore, be viewed as a success in technology transfer.

3.2 CAD Model Acquisition from Multiple Range Images

Automating the acquisition of Computer-Aided-Design (CAD) surface or solid models from laser scan data has been identified as one of the major goals in the field of computer vision. As CAD models become more central to parts design and manufacture, the ability to automatically generate these models from existing objects becomes paramount. There are still parts which are best designed using the tools of model makers, in materials such as clay or wood. It has been said that everyone would be using CAD systems if they were "as comfortable and easy to use as foam, clay, and pine". As long as this state of affairs continues, there will be parts for which there are no CAD data. Without CAD data, it is not possible to use rapid prototyping systems to produce additional models, nor is it possible to benefit from any of the advanced analysis, manufacturing, and process planning capabilities of today's CAD systems. Other applications in which 3-D solid or surface data must be acquired from physical models or prototypes include model making, inspection and quality assurance, and reverse engineering.

In our work we describe an approach to automated model acquisition that combines work in range data acquisition, segmentation and polyhedral model construction to address some of the problems cited above [25]. We motivate use of Binary Space Partitioning Trees (BSP trees) as an intermediate data structure that can easily be derived from low-level scanned range data, and from which multiple views can be efficiently merged into a single B-rep description from which a CAD model may be derived. The BSP tree represents volumes by partitioning space with planes, and therefore is limited in that it may only represent polyhedra. It does, however, have other attributes which make it a very attractive primitive for modeling 3-D objects, including both robustness and the existence of efficient algorithms for set operations.

Our system works in a cycle consisting of three phases: acquisition and preprocessing, segmentation and model generation, and integration. In the first phase, a laser rangefinder is positioned by a robot arm to acquire a range image to which low-level processing is applied. The position and orientation (or *viewpoint*) that the arm holds the rangefinder in is known beforehand or, in future work, will be computed dynamically. In the second phase, the image is segmented into regions and a BSP tree model is built of that single image. Since the range image is from a particular viewpoint, this model will not be closed on the side opposite the rangefinder. In the third phase, the model is integrated with the models created from other views using boolean operations on the BSP trees. The final model is produced when the BSP tree model of each view has been integrated. Figure 4 shows laser range data for 3 views of an object, figure 5 shows the segmentation of the data, and figure 6 shows the models built from each view with their occlusion volumes. Each time a single-image model is generated it is added to the composite model using boolean operations. These models can then be merged into the composite model.

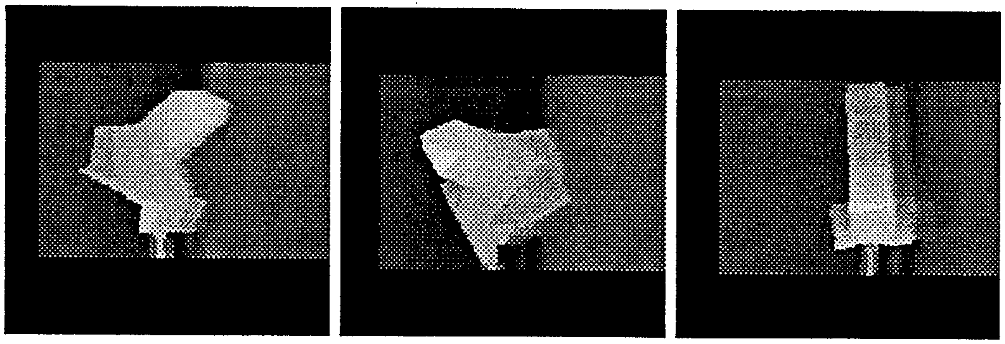


Figure 4: Real range data of part taken from three different views.

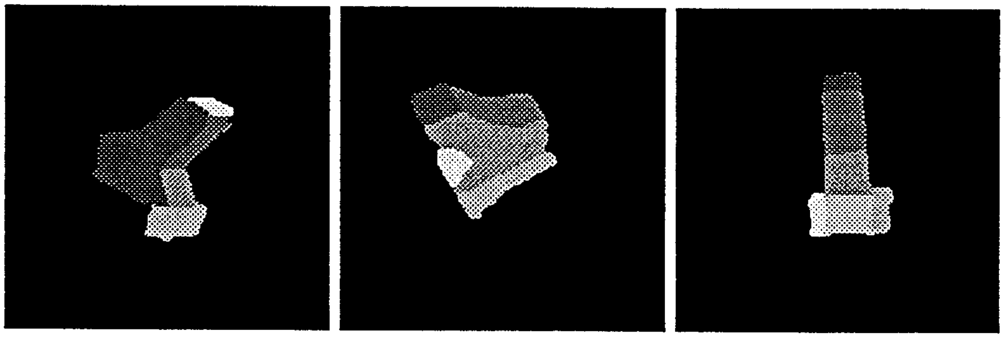


Figure 5: Segmentations of the real range data.

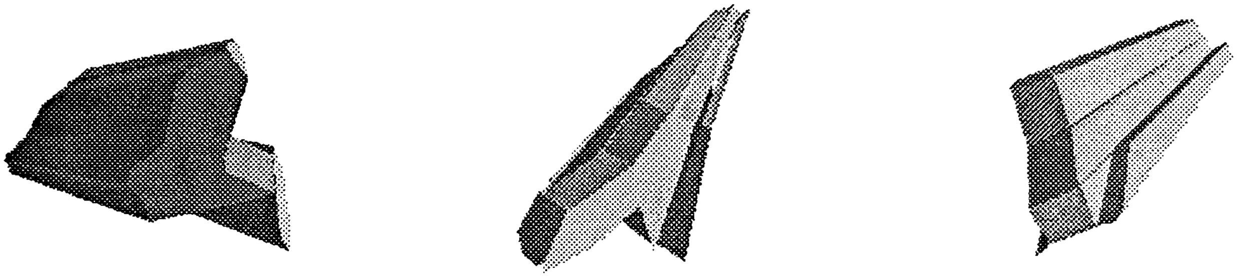


Figure 6: BSP tree models of the segmented faces and occlusion volumes.

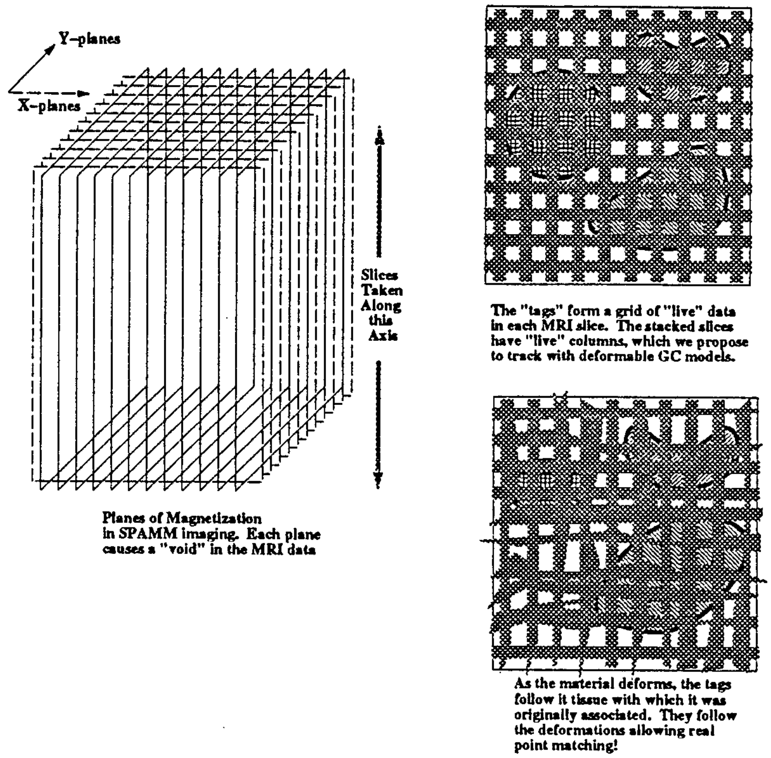
3.3 Deformable Object Modeling

We have been developing software for deformable object modeling based on deformations from underlying generalized cylinders (GC's). The work is currently being pursed in the context of medical imaging, but the techniques can also be applied to tracking people, thermal front tracking (atmospheric and oceanographic), cloud/plume tracking, seismic analysis and may also be used on non-deformable objects. This work is building on our previous work on deformable GCs and super-quadratics, and has continued on two fronts: model extensions with implementation-al/experimental aspects and theoretical reformulation. Final results from our related work on symmetry recently appeared, [8] and the final results from our generalized cylinder recovery/modeling will appear soon, [9].

Our theoretical work has looked at reformulating deformable model recovery in terms of robust statistical M-estimators. These ideas were presented at the National Science Foundation (NSF)/Advanced Research Projects Agency (ARPA) Workshop on Representations for 3-D vision, [7]. (Related material was presented in a paper in the proceedings of the ARPA IUW workshop and in a seminar at the University of Washington). The basic idea was to recast the "forces" of physically-motivated deformable modeling as the weighting functions of robust M-estimators. This approach naturally suggested improvements to the process which were incorporated into our cardiac modeling work, and have been submitted for conference publication. While other researchers have related deformable fitting to statistical estimation, our approach relates it to Robust Statistics. This reformulation has allowed us to provide solid justification for commonly used "ad hoc" techniques as well as suggesting new variations. For example, most of the ad-hoc, but basically successful techniques of deformable object fitting, cannot be justified in terms of pure model/data uncertainty, but *can be justified in terms of robust estimation and outlier processes.*

The experimental work is novel for two important reasons. The models are true temporally deforming volumes. They are not just surface "shells", as in almost all previous work, but rather real volumetric solids. Secondly, we have extended the technique to handle data constraints that are projections of the actual constraints. This is important if the ideas are to be extended to traditional camera imagery.

Our current project is using SPAMM (SPAtial Modulation of Magnetization) data to build a 4-D cardiac model. This is a continuing part of our collaboration with Siemens Research. The SPAMM data provides the opportunity for direct 2-D material correspondences, projections of the true 3-D motion of the heart, see figure 7. The SPAMM tags are too short lived to provide for periodic tracking, and we halted our periodic modeling efforts and concentrated on building a full 4D cardiac model using a combination of contour and SPAMM data. We call the new model the Hybrid Volumetric Ventriculoid (HVV), a hybrid deformable model designed for recovery of the left ventricle. The HVV is composed of an implicit parametric global volume component and a set of volumetric finite elements which express both rest-state of the model and local deformations away from this underlying shape, see 8. The global model is (currently) a thick-walled super-ellipse, though we are also exploring a GC-based version. The model has a rest-state which is defined as parametric offsets



Planes of Magnetization in SPAMM imaging. Each plane causes a "void" in the MRI data

The "tags" form a grid of "live" data in each MRI slice. The stacked slices have "live" columns, which we propose to track with deformable GC models.

As the material deforms, the tags follow it tissue with which it was originally associated. They follow the deformations allowing real point matching!

Figure 7: SPAMM data imparts a physical pattern of magnetization on the heart tissue. This increase our ability to track tissue as the heart deforms. However tracking the pattern (tags) does not constrain the motion through the imaging plane, i.e. the Z-axis.

from the global model, so the "rest-state" is a GC. This rest state is, in a statistical sense, our prior model; all deformation penalties are measured from the rest-state. If there is no data, the model assumes the rest-state. Because our rest-state is much closer to underlying object than a regular super-ellipsoid, we have a more meaningful prior distribution and our FEM deformations are kept small. A true FEM model is used—uncertainty propagates using the FEM shape functions. Multiple-FEM layers are permitted between slices and between endocardium and epicardium.

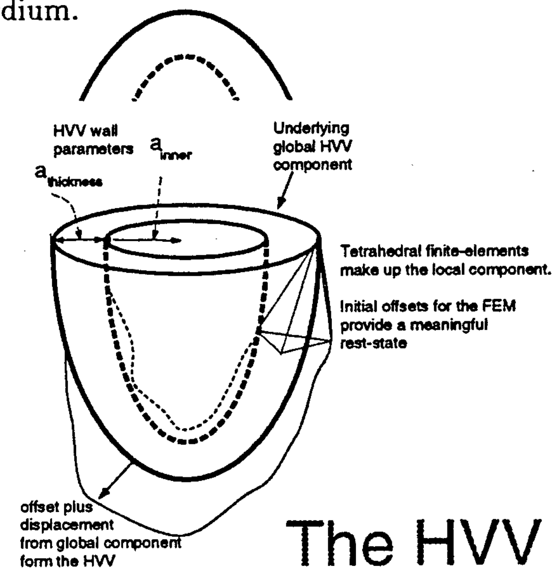

HVV wall parameters
a_{inner}
$a_{thickness}$
Underlying global HVV component
Tetrahedral finite-elements make up the local component.
Initial offsets for the FEM provide a meaningful rest-state
offset plus displacement from global component form the HVV
The HVV

Figure 8: The Hybrid Volumetric Ventriculoid (HVV) is a deformable solid which can be used to track/recover the left ventricle of the heart.

Recover of the HVV is comprised of four stages.

1. HVVs are fit to contour data from the different phases (time slices) in the cardiac cycle.

   The model offsets are set to the model displacement values, and the model displacements nullified.

   The result is a set of rest states resembling the cardiac geometry at different phases. (In future work this will come from an atlas).

2. The contour defined HVV models from t>1 are deformed, according to the 2-D displacement information from multiple orthogonal SPAMM acquisitions, to resemble the initial state (t=1). This "registration" of multiple orthogonal sets of 2-D displacements allows us to infer *3-D* displacements.

   During this deformable fit, the SPAMM intersections are constrained only to match in their in-plane coordinates. The thru-plane coordinate is free to change in response to the influences of alternate view information.

3. The 3-D movement of the tags garnered from stage 2 is reversed to deform the initial rest state forward in time. The result is full 3-D displacements in each frame, see 9

4. The contour data can now be reintroduced, thereby providing more detailed information about the shape of the outer boundary of the model. This stage fits to the combination of the displacements from stage 3 and the contour data. (Currently being implemented)

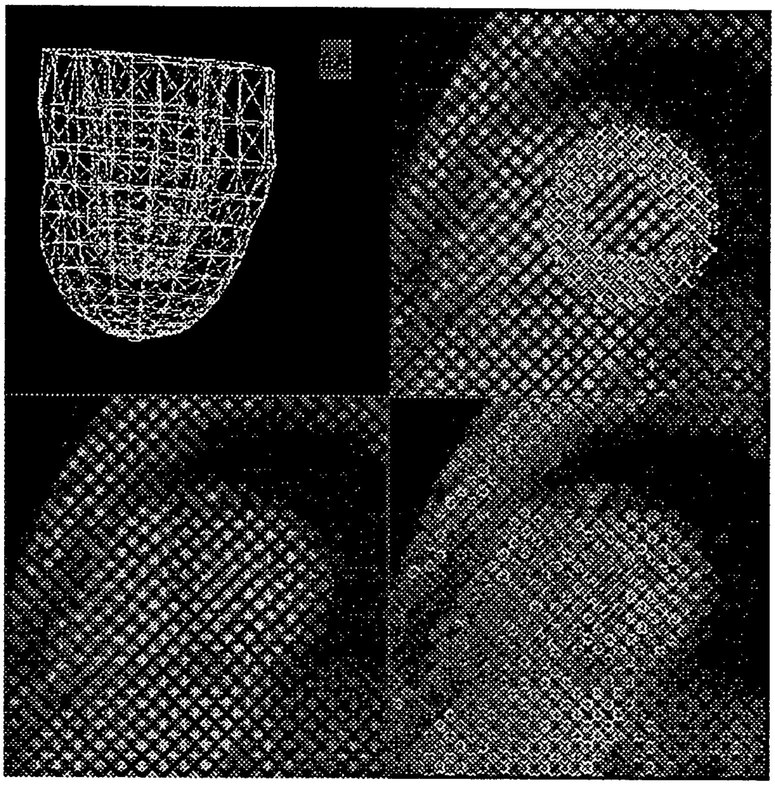


Figure 9: Results of HVV fitting. Lower right quadrant shows one slice of SPAMM data, lower right shows some of the forces from that slice. Upper right quadrant shows the tracked contours and intersection points from that slice. The upper left shows the inner and outer walls of the final HVV model associated with the same time as the slice.

Compared to previous work our approach has the following advantages:

+ Unlike previous work (e.g. Axel & Young's approach, [5]), the HVV is a hybrid model, and therefore is able to provide concise descriptions of overall shape and movement supporting comparison of an left ventricle (LV) under study with a "normal" population to detect and classify abnormalities.

+ A local spline-like component describes fine detail, while the global component gives us the ability to fit more rapidly, and with increased topological stability, compared to purely local models.

+ The HVV recovery paradigm also extends the state-of-the-art to directly include myocardial contour information, and mixes tracking and fitting. Axel & Young make use of the SPAMM grid intersections solely in fitting their model; Park, Metaxas, and Young, use the data from Axel & Young.

+ Unlike the super-elliptic hybrid model proposed by Park, Metaxas, and Young, ([24]) the HVV is a **volumetric** model. Thus, the SPAMM displacements can act directly on the model to deform it.

+ The novel use of offsets from the global component allow the HVV to closely resemble the LV, even in the absence of local component activity, i.e we have **a realistic prior model.**

A summary of the deformable modeling work, including a movie of the results of the fitting process, will be available at URL http://www.eecs.lehigh.edu/boult/DEFORM after June 30 1995. Note the most recent work incorporates some the ideas suggested by our theoretical work which improved the quality and speed of fitting.

In addition to the above publications, Dr. Boult helped organize and run the NSF/ARPA workshop on representations for 3-D vision and prepare the report, which will be published as a book in the Springer Verlag series Lecture Notes in Computer Science. The basic report and paper abstracts will be available soon on the World-Wide Web (WWW).

3.4 Shape from Darkness

One method for determining object shape from imagery is to study the nature of the shadows that an object casts, both on the ground and on itself. We have previously developed and patented a method for taking a series of images of an object under various sharp illumination conditions, extracting the shadows, generating constraint relationships among them, and reconstructing the object shape from these relationships. The method works perfectly on perfect data, but on real-world noisy data it usually fails to converge to a solution. We have analyzed the several sources of error that arise, and have approached the problem from a different mathematical perspective. This has resulted in a newer and very robust demonstration system.

The algorithmic foundations of this system have been completely recast as a problem in linear programming, with a net speedup of over a factor of ten, and with an error rate effectively of zero. All other observations and behaviors remain valid; in particular, finding the "best" consistent set of imagery remains NP-hard. It is anticipated it will be defended as a thesis in September 1995.

Algorithmic and performance improvements include:

1) Developed an algorithm to detect the existence of conflicting shadow constraints in $O(N^3)$ time; it is roughly equivalent to Early's algorithm in method and data structure.

2) Discovered a way of representing the shifting of shadows as a linear programming constraint problem. In particular, instead of trying to alter the incoming image to minimize reconstruction error, the image is instead warped algorithmically, by real (not integer) amounts. These "shift equations" can be shown to be consistent, or in conflict, and by the addition of linear programming goal variable techniques, can be made to minimize the amount of image distortion necessary while simultaneously yielding a consistent image set. Also necessary for the method to work were additional constraints limiting the shifts to those that preserve image topology: the image cannot be "torn". Thus, instead of an A* heuristic search for the best modification to incoming imagery, the problem is now cast as a linear programming simplex method for the least amount of image warp.

3) Discovered, taxonomized, and demonstrated a way to "decouple" those constraints which cannot be handled by image shift; these errors are exceedingly rare in practice, and basically occur only at extremely shallow illumination angles.

3) Discovered a method for compressing incoming shadow data so that the linear programming method need not operate on the image itself, but on a simpler set of data which is approximately equivalent to shadow edge data. As the number of images taken at different light illumination angles increase, these two data sets approach each other.

4) Formally proved many of the fundamental underlying results of the Shape from Darkness method. For example, formally proved that errors in incoming data can be detected simply by observing the reconstruction behavior at one critical pixel, namely, the global maximum height pixel. Also proved that the shadow edge equivalent set of imagery is in fact sufficient for a full surface reconstruction, and demonstrated how the full surface can be recovered from it.

5) Implemented the method, and validated its performance on imagery captured from objects made of the widely varying materials of wood, styrofoam, and aluminum foil, without the use of any reflectance information or calibration.

6) Designed and constructed a novel apparatus for mounting objects on a robot arm so that the arm can present the object to different angles of illumination; this apparatus allows imagery to be illuminated from light directions that vary over the surface of the Gaussian hemisphere.

4 ROBOTIC VISION

4.1 Visual Control of Grasping

In most manufacturing tasks, it is necessary to have the ability to move parts together in useful configurations to make an assembly process more efficient. For example, we might require a robot to grasp a part off the assembly line and insert it into some fixture. In order to perform the grasping portion of this task, the robot must be able to locate the part, move its gripper to the vicinity of the part, locate the best points to grasp the part,

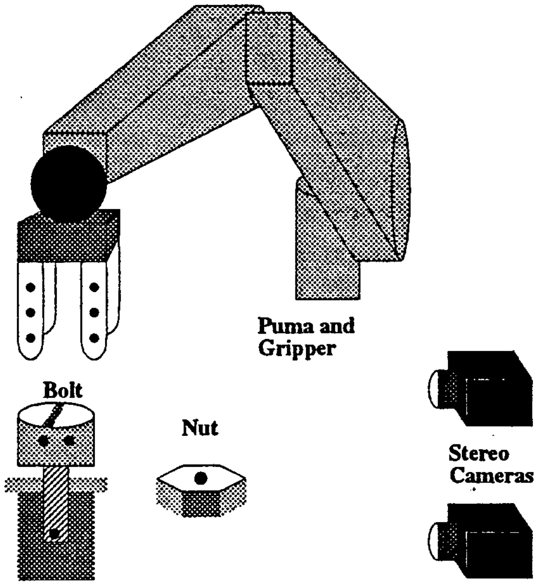


Figure 10: Experimental system used to test visual control of grasping and manipulation.

move the fingers of the gripper to those grasping points, and finally, verify that the object is stably supported. Usually these kinds of tasks are performed by blind robots which use world coordinates, jigs and other devices to remove the need for the robot to visually find the 3-D location of the object. Systems built using this kind of robot control usually require large start-up costs in pre-production measurement, setup and testing. These systems also exhibit inflexible, brittle qualities. If a small design change is made in how the product is manufactured, the cost of replanning the robot assembly line, which may include extensive retooling, rejigging and a total revision of the robot control strategy, can be prohibitively expensive. The research we have completed addresses some of the problems associated with vision-based robot control [33, 32, 34, 31]. The system, shown in figure 10, shows the major components of the system: a robot, a gripper, and two fixed stereo cameras. The vision system is able to track multiple moving targets at real time rates. In our system, each moving target is a fiducial mark (a black dot on a white background) which can be attached to a robot or other objects in the environment. The upper bound of the number of targets trackable at any one time is 255. Each finger of the robotic hand has 4 fiducial marks, and each object to be manipulated also has 1 or 2 fiducial marks. The tracker uses intensity thresholds to segment the fiducial marks from the background.

Using this method, we have shown how vision can be used to direct the movement of a Puma robot and the movement of a two-fingered gripper system. In both cases, the robot systems were entirely unaware of the environment surrounding them (both systems were sensorless.) Under visual control, a Puma robot aligned and inserted a bolt into a nut, and the attached gripper grasped and tightened a bolt. This system performed tasks which would have been difficult to accomplish without the spatial understanding provided by vision.

The key to our solution lies in the development of primitive visual control operations and the application of a hierarchical approach to the visual control problem. The primitive visual control operations allowed us to examine many of the real time problems associated with visual control in concise, manageable units. By decomposing a complex manipulation into a series of these operations, we removed much of the complexity associated with creating a visual control system. Also, by developing the system in a modular fashion, we were able to readily reuse many of the primitive operations and complex tasks to solve many more complex problems. Not only was it not necessary to add sensors to the fingers, but the results have shown that visual control is capable of performing many of the operations necessary for elementary manipulation (information which was previously obtained using other sensors.)

4.2 Dynamic Sensor Planning

In this project, we have been researching methods for the automatic computation of viewpoints for monitoring objects and features in an active robot work-cell. We call this "Dynamic Sensor Planning." The static sensor planning problem has received much attention lately. Most research has focused on the computation of sets of positions, orientations, and optical settings for a camera (and, in some cases, for light sources) which will give satisfactory views of certain objects in a known scene. Each researcher has defined the phrase "satisfactory view" in his own terms, but the constraints most often considered are magnification (or resolution), focus, field-of-view, and occlusion.

We have been working on extending our Machine Vision Planning (MVP) system [26, 28, 27] to function in an environment in which objects are moving. As a specific example, consider a robot which has to deposit a bead of glue on a part prior to using that part in an assembly. A dynamic sensor planning system can be used to compute either stationary viewpoints or camera trajectories which maintain an unobstructed, well focused view of the target area on the part to monitor the gluing process.

The basic setup includes two Puma 560 arms, able to operate in a work-cell, and a gantry robot, used for moving the camera through a computed trajectory. Our approach to Dynamic Sensor Planning has been based on temporal intervals, in which the task is broken down into intervals, each of which is to be monitored by a single viewpoint. To solve the occlusion problem, the system computes the volumes swept by all moving objects during this interval and, using the algorithms developed as part of MVP, computes viewpoints which avoid occlusion by these swept volumes. Such viewpoints are valid for the entire time interval. By similarly examining a number of time intervals, we break the Dynamic Sensor Planning problem down into a series of static subproblems. We have been using MVP to compute the viewpoints for each of the static subproblems, although part of the current work is focused on better methods of computing viewpoints for the static subproblems.

The general idea behind our approach to dynamic sensor planning is best described by the following Temporal Interval Search algorithm [3]:

1. Compute the volumes swept by all moving objects during the task interval T.

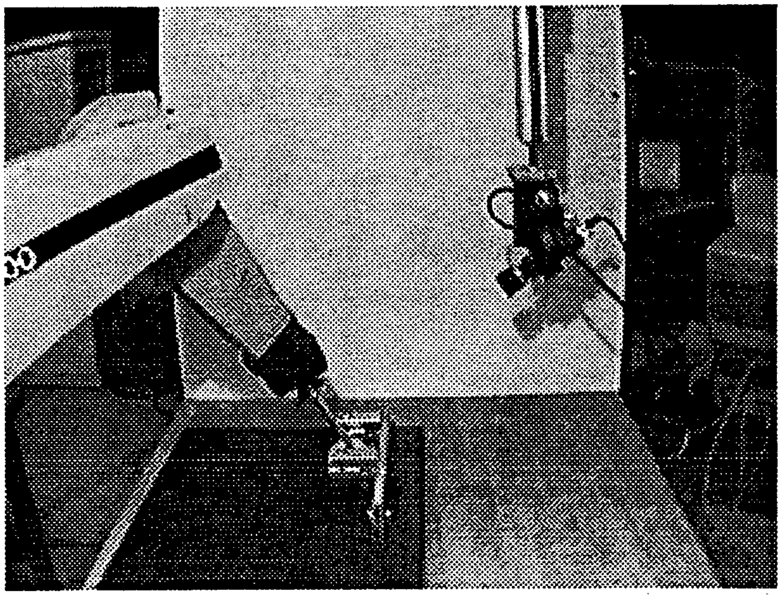


Figure 11: Overview of the Experimental Setup

2. Use MVP to compute a valid, unoccluded viewpoint using these swept volumes as if they were actual objects.

3. If MVP can successfully find a viewpoint, use this viewpoint for the entire time interval T.

4. If no such viewpoint is obtainable, divide the time interval in half yielding $T_1 = [t_0, t_{n/2}]$. Go back to step 1 using interval T_1.

5. If the entire time interval T has been planned, we are finished. If not, go to step 1 using the remaining portion of the the original interval T.

Thus, a swept volume algorithm useful for this research must produce results fairly quickly (since it may be called often), and must produce polyhedral models of these volumes. Therefore, we have embarked upon the present research task of robustly and efficiently computing polyhedral approximations to the volumes swept by moving polyhedral objects. A necessary step of computing these volumes is computing the volume swept by a moving *polygon.*

We have developed an algorithm for computing these polyhedral approximations to these volumes [4]. We create a set of polygonal faces which are a superset of the boundary of the actual swept volume, by computing and approximating each of the 3 types of faces for each of the polygons in the polyhedron under motion. Then, we compute the arrangement of this set of faces, and traverse the outer boundary of this arrangement. This bounds our approximation to the swept volume.

In order to demonstrate the usefulness and practicality of the sensor planning system as a whole, we will be running experiments in our robot work-cell. In these experiments, surveillance points and intervals will be computed by the dynamic sensor planning system. To realize these computed viewpoints, we have been constructing a sensor-positioning robot. This 5 degree-of-freedom Cartesian robot, having a work-space of roughly 1000 ft^3, carries a camera in calibrated hand/eye configuration, and is pictured in figure 11. It can accurately position the camera in and around our robot work-cell, thereby monitoring objects under manipulation by our two Puma 560's. The hardware and controller software have been implemented, and as soon as the hand/eye system is calibrated, we shall continue with sensor planning and placement experiments.

5 INTEGRATION OF VISION WITH OTHER MODALITIES

5.1 Language Description of Visual Images

People don't use everything they see. When they describe something of importance, they have several mental skills that allow them to summarize what is important in the observed scene, and to communicate that importance effectively. We have studied this problem and built a demonstration system that show how people understand significant spatial positions, orientations, and relationship of objects and how they express those concepts in English. The system works in two domains: it can talk about where kidney stones are in radiographs, and it can give directions to a building in a theme park such as Epcot center. This system was analyzed and improved in numerous ways, and then validated against the performance of people. The system and its documentation was defended as a thesis [2] in January 1995, and has produced a workshop paper submission [1]. A journal article incorporating these results is in preparation.

Improvements included:

1) A complete rewriting of the notation for the concepts involved, mapping the concept of spatial relationship into first-order predicate calculus, and adopting predicate calculus notation for exposition.

2) Debugging the concept of minimization of spatial description, by adopting a form of three-valued logic to represent the three important classes of spatial concepts. First, namely, those relationships between a reference object and the to-be-described figure object, second, those relationships between a reference object and any of the distractor objects in the same visual context, and third, those relationships that are completely unused. The Quine-McClusky minimization technique was employed (which is of exponential order), but a heuristic variation on the method was also discovered that has a much shorter expected running time (of polynomial order). The method was extended to cover the concept of fuzzy minimization, as well, although crisp minimization gave adequate results and ran much faster.

3) Defined the role of adverbials in modifying spatial descriptions, such as "very" near, etc., and algorithmically implemented it as a type of post-processing approximation to fuzzy

minimization. In short, the system of selecting an unambiguous deep spatial relationship now consists of three processing steps: first, the fuzzy individual relationships derived from the image data are represented as crisp logical predicates; second, the intended relationship is selected through a three-valued logic minimization criterion; third, the fuzz is put back in through the use of modifying adverbials.

4) Made algorithmic the meaning of "descriptiveness" by defining it to be the probability that a user would properly select the intended figure object over any of the surrounding distractors. Since this depended on a model of user error, the performance of humans was measured in a series of tests that captured how well people agreed in the use of simple prepositions such as "right" or "near". It was found, for example, that objects having visual measurements of very low or very high image nearness, were also so described reliably by humans, however the vast middle ground of relationships had extremely high variance. Since it was difficult to obtain closed form solutions to the convolution of a user's model of the figure's spatial description with that of the distractors, the definition of "descriptiveness" is defined stochastically by dynamic simulation.

5) Defined and implemented spatial inference, by noting that humans tend to pick as descriptive intermediate objects, those objects which imply the most other spatial relationships. Thus, no kidney stone is ever described as "above the pelvis" since everything in a kidney radiograph is above the pelvis; instead, phrases like "in the calyx" are selected because they imply, among other relationships, "in the kidney" and "above the pelvis". Thus, a "good" description is one that implies the most descriptions.

6) Developed and documented the existence of a novice-expert continuum in spatial description. In this, the novice describes all relationships with respect to very few "obvious" figure objects; in the extreme, there are no figure objects at all, and all objects are described with respect to image boundaries. Conversely, for an expert, nearly all objects in an image are "obvious"; the task is instead one of filtration of relationships, rather than creation. Middle grounds are possible, and the two methods of relationship creation and filtration are complementary.

7) Analyzed and made algorithmic the concepts of comparative and superlative spatial relationships, noting that extreme measures are often easily and correctly determined by humans.

Verification and validation included:

1) Experiments with sixteen people who were given maps of the Epcot center, and descriptions of ten objects in the map, all of which were described using only spatial relationships. Agreement was high, with most of the subjects correctly identifying the ten objects, and only some confusion, which was systemic.

2) The validation established a systemic cause of interpersonal error, namely, that people define relationships such as "right" as either "strictly right" (roughly speaking, due east), or "generally right" (roughly speaking, anywhere from northeast to southeast). People are self-consistent over time, but definitions between people are highly variable.

5.2 Visual Hand Gesture Recognition

Talking with one's hands is not only natural, it can be useful as well. We are developing a system that replaces a personal computer's mouse with a camera instead. The pointing and clicking of the mouse can then be replaced by gestures and movements of the hand itself. The system tracks the hand and interprets a primitive "sign language", driving a menu selection system that appears on the screen. Work on this system was incremental, and has lead to the acceptance of a workshop paper [11] . Research included the following experiments and results.

1) Training data for the neural network front-end that classifies (static) gestures into one of four categories, was collected, and the network trained, with higher than 80 percent accuracy in actual tests. The network was optimized on a workstation. Some time was spent to automate calibration, particularly under different light conditions.

2) Image tracking of the centroid of the segmented hand was optimized for stand-alone performance on a DSP board in the workstation; complete segmentation and production of (x,y,scale) triples of the hand now runs in excess of 7 Hz. Gestures were shown to have idiosyncratic signatures in this space.

3) Noticed, documented, and exploited in algorithms, that the signal equivalent of "gesture begin", "gesture end", "qualifier begin", and "qualifier end", are apparent simply in the (x,y,scale) triples, with local extrema in y, particularly, indicating gestural punctuation. The y dimension appears to dominate, in part because it is this dimension in which the most work is done by the human arm in resisting gravity. These "haptic prosodics" clues are critical for triggering a full image capture for diagnosis by the neural net. Thus, the DSP board detects event boundaries, and the workstation proper analyzes and quantifies gestural content. The full system now tracks and interprets, although not fully automatic.

6 LICENSEES OF THE SLAM PACKAGE: 94-95

Department of Computer Science, University of Rochester;
Department of Computer Science, University of Virginia;
Department of Electrical and Computer Engineering, Lehigh University;
Department of Electrical Engineering and Computer Science, Lehigh University;
Department of Electrical and Computer Engineering, Drexel University;
The Robotics Institute, Carnegie Mellon University;
Department of Computer Science, Michigan State University;
Department of Electrical and Computer Engineering, Pennsylvania State University;
Artificial Intelligence Laboratory, Massachusetts Institute of Technology;
Department of Electrical and Computer Engineering, University of California, San Diego;
Courant Institute, New York University;
Department of Mechanical and Environmental Informatics, Tokyo Institute of Technology;
Department of Mechanical Engineering, Osaka University;
Department of Computer Science, University of Massachusetts;
Information Science Research Laboratory, NTT Basic Research Laboratory;
Instituto de Cibernetica, Barcelona, Spain;
Department of Computer Science, University of Genoa, Italy;
Computer Vision Laboratory, LIFIA, France;
Center for Applied Computer Science, G.Fa.I.e.V., Germany;
Morgan Stanley Research Division, New York;
Quality Control Department, Pressco Limited, Ohio;
Cambridge Research Laboratories, DEC, Cambridge

7 BIBLIOGRAPHY AND RECENT PUBLICATIONS 94-95

[1] A. Abella, J. Starren, and J. R. Kender. Automated natural language description of radiographs. In *Proceedings of the 19th Symposium on Computer Applications in Medical Care*, October 1995.

[2] Alicia Abella. *From Imagery to Salience: Locative Expressions in Context.* PhD thesis, Department of Computer Science, Columbia University, 1995.

[3] S. Abrams, P. K. Allen, and K. A. Tarabanis. Dynamic sensor planning. In *Proceedings DARPA 1993 Image Understanding Workshop*, Washington, DC, April 1993.

[4] Steven Abrams and Peter K. Allen. Swept volumes and their use in viewpoint computation in robot work-cells. In *Proceedings IEEE 1995 International Symposium on Assembly and Task Planning*, Pittsburgh, PA, August 1995.

[5] L. Axel and L. Dougherty. Heart wall motion: Improved method of spatial modulation of magnetization for mr imaging. *Radiology*, 172:349–350, 1989.

[6] D. Bhat and S. K. Nayar. Stereo in the presence of specular reflection. *International Conference on Computer Vision*, pages 1086–1092, 1995.

[7] T.E. Boult, S.D. Fenster, and T. O'Donnell. *Physics in a Fantasy World vs. Robust Statistical Estimation.* Springer-Verlag, Heidelberg and New York, 1995. Proceedings of the NSF Workshop on 3D representation, to appear.

[8] A.D. Gross and T.E. Boult. Analyzing skewed symmetries. *International Journal of Computer Vision*, Nov 1994.

[9] A.D. Gross and T.E. Boult. Understanding straight homogeneous generalized cylinders: A case study. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 1995. To appear.

[10] J. R. Kender and R. Kjeldsen. On seeing spaghetti: A novel self-adjusting seven parameter hough space for analyzing flexible extruded objects. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, February 1995.

[11] R. Kjeldsen and J. R. Kender. Visual hand gesture recognition for window system control. In *Proceedings of the IEEE International Workshop on Automatic Face- and Gesture-Recognition*, Zurich, June 1995.

[12] Paul Michelman and Peter Allen. Forming complex dextrous manipulations from task primitives. In *1994 IEEE International Conference on Robotics & Automation*, San Diego, May 1994.

[13] S. Mukherjee and S. K. Nayar. Automatic generation of grbf networks for visual learning. *International Conference on Computer Vision*, pages 794–800, 1995.

[14] H. Murase and S. K. Nayar. Learning object models from appearance. In *Proc. of AAAI*, Washigton, July 1993.

[15] H. Murase and S. K. Nayar. Illumination planning for object recognition in parametric eigenspaces. In *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition*, June 1994. Outstanding Paper Award.

[16] H. Murase and S. K. Nayar. Illumination planning for object recognition in structured enviroments. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, December 1994.

[17] H. Murase and S. K. Nayar. Visual learning and recognition of 3-d objects from appearance. *International Journal of Computer Vision*, April 1995.

[18] S. K. Nayar, X. Fang, and T. E. Boult. Separation of reflection components using color and polarization. *International Journal of Computer Vision*, 1996 (in press).

[19] S. K. Nayar and M. Oren. Visual appearance of matte surfaces. *SCIENCE*, 267:1153–1156, February 1995.

[20] M. Oren and S. K. Nayar. Comprehensive model for diffuse reflection. *Proceedings of SIGGRAPH 94*, July 1994.

[21] M. Oren and S. K. Nayar. Generalization of the lambertian model and implication for machine vision. *International Journal of Computer Vision*, April 1995.

[22] M. Oren and S. K. Nayar. A theory of specular surface geometry. *International Conference on Computer Vision*, pages 740–747, 1995.

[23] I. P. Park and J. R. Kender. Topological direction–giving and visual navigation in large enviroments. *Artificial Intelligence Journal, to appear, Special Issue on Computer Vision*, 1995.

[24] J. Park, D. Metaxas, and A. Young. Deformable models with parameter functions: Application to heart-wall modeling. In *Proceedings of the IEEE CVPR, Seattle , Washington*, pages 437–442, 1994.

[25] Michael Reed, Peter K. Allen, and Steven Abrams. CAD model acquistion using BSP trees. In *IROS International Conference on Intelligent Robots and Systems*, pages 335–339, August 1995.

[26] K. Tarabanis, Roger Tsai, and Peter K. Allen. Analytical characterization of the feature detectability constraints of resolution, focus and field-of-view for vision sensor planning. *Computer Vision, Graphics, and Image Processing*, 59(3):340–358, May 1994.

[27] Konstantinos Tarabanis, Peter K. Allen, and Roger Y. Tsai. A survey of sensor planning in computer vision. *IEEE Transactions on Robotics and Automation*, 11(1), February 1995.

[28] Konstantinos Tarabanis, Roger Y. Tsai, and Peter K. Allen. The MVP sensor planning system for robotic vision tasks. *IEEE Transactions on Robotics and Automation*, 11(1), February 1995.

[29] Alex Timcenko and Peter Allen. Probability–driven motion planning for mobile robots. In *1994 IEEE International Conference on Robotics & Automation*, San Diego, May 1994.

[30] B. Yoshimi and P. K. Allen. Visual control of grasping and manipulation. In *Proc. ARPA 1994 Image Understanding Workshop*, pages 1151–1158, November 1994.

[31] Billibon Yoshimi. *Visual Control of Robotics Tasks.* PhD thesis, Dept.of Computer Science, Columbia University, 1995.

[32] Billibon Yoshimi and P. K. Allen. Visual control of grasping and manipulation tasks. In *IEEE Multi-Sensor Fusion '94*, October 1994.

[33] Billibon Yoshimi and Peter Allen. Active uncalibrated visual servoing. *IEEE Transactions on Robotics and Automation*, 11(5):516–521, August 1995.

[34] Billibon Yoshimi and Peter. K. Allen. Active uncalibrated visual servoing. In *IEEE International Conference on Robotics and Automation*, pages 156–161, San Diego, 1994.